Can accelerometry data improve estimates of heart rate variability from wrist pulse PPG sensors?*

Maciej Kos, Xuan Li, Iman Khaghani-Far, Christine M. Gordon, Misha Pavel, Senior *Member, IEEE,* and Holly B. Jimison *Member, IEEE*

Abstract— **A key prerequisite for precision medicine is the ability to assess metrics of human behavior objectively, unobtrusively and continuously. This capability serves as a framework for the optimization of tailored, just-in-time precision health interventions. Mobile unobtrusive physiological sensors, an important prerequisite for realizing this vision, show promise in implementing this quality of physiological data collection. However, first we must trust the collected data. In this paper, we present a novel approach to improving heart rate estimates from wrist pulse photoplethysmography (PPG) sensors. We also discuss the impact of sensor movement on the veracity of collected heart rate data.**

I. Introduction

Several countries have developed groundbreaking initiatives to accelerate new advances in Precision Medicine, [1, 2] urging healthcare professionals, researchers, and policymakers to transition from population-level, reactive sick-care to personalized, proactive prevention. With the recent advent of widely available wearable sensors, the possibility of proactively assessing individuals' health in natural settings in a continuous and unobtrusive manner is within our reach. By monitoring critical physiological markers of cardiovascular, autonomic, and mental health, such as heart rate variability and electrodermal activity, we can enable the delivery of just-in-time adaptive interventions [3] and help to optimize individuals' health, thus making precision medicine a reality. Among many factors, our ability to do so hinges upon understanding the quality of physiological data obtained from wearable sensors. This paper moves us toward this goal by presenting results of a careful laboratory experiment assessing the quality of data obtained from two wearable wrist sensors (Microsoft Band 2 and Empatica E4), specifically considering the accuracy in measuring heart rate as a precursor to estimating the important indicator of heart rate variability.

The degree of changes in timing between heartbeats – heart rate variability (HRV) – is an informative metric in a surprisingly broad range of contexts. Long-term HRV (captured over 24 hours) can be indicative of autonomic dysfunction and cardiovascular health[4]. Even short-term HRV measures (e.g., standard deviation of beat-to-beat intervals captured over 2 to 5 minutes) reflect a wide range of issues, including cachexia,[5] hypertension,[6] stress,[7] regulation of emotion, and depression [8]. Given the enormous impact of cardiovascular disease, stress, and mood disorders on population health outcomes, continuous monitoring of HRV at the individual level may prove to be an immensely important aspect of clinical care [9] [10].

Analysis of HRV relies mostly on linear methods. (Nonlinear methods, while promising [11], are not used as often.) Linear methods include 1) time-domain measures, e.g., standard measures of central tendency and spread computed over a time interval, ignoring the order of observations, and 2) frequency domain measures providing information about a relative proportion of high and low frequency signals in a given time frame [12].

Time-domain indices, such as the standard deviation of all beat-to-beat intervals or square root of the mean of the squares of the differences between adjacent beat-to-beat intervals, are well-established biomarkers of cardiovascular health. Frequency-domain measures have been shown to represent the activity of the autonomic nervous system. In general, low-frequency (LF) is modulated by both sympathetic and parasympathetic systems, and high frequency (HF) represents parasympathetic activity only [13]. The ratio of LF to HF is representative of the sympathico-vagal balance [12].

Accurate continuous measurement of HR in natural settings is non-trivial. The majority of research relies on HR captured using electrocardiography (ECG). This approach, while highly accurate, is impractical for long-term monitoring because it requires electrodes to be attached to the patient's chest, which is inconvenient and may interfere with day-to-day activities (e.g., showering or sleeping) and may cause skin irritation if used over several days. Alternatively, HR can be derived from blood volume changes measured at extremities such as the wrist, finger or earlobe using pulse photoplethysmography (PPG). Laboratory studies using PPG data usually rely on measurements obtained from subjects' finger or earlobe. While pulse rate variability obtained this way is considered to be an accurate estimate of heart rate variability [14] when subjects are not moving [15], this data capture approach is too inconvenient for continuous, long-term monitoring.

A more promising mode of data collection is one using relatively unobtrusive wrist PPG sensors embedded in wrist watch-likedevices. With a few exceptions, validity studies of wearable sensors largely focus on fitness-related measures (e.g., number of steps, energy expenditure) and not HRV [16]. Existing research that does focus on HR measures suggests

*Research supported by the National Institute of Disability, Independent Living, and Rehabilitation Research (Grant # 90RE5023-01-00) and the National Institute of Nursing Research (Grant # P20NR015320).

M. Kos is with the Consortium on Technology for Proactive Care, Northeastern University, Boston, MA 02115 USA (phone: 617-373-2381; e-mail: mkos@ccs.neu.edu).

X. Li, I. Khaghani-Far, C. M. Gordon, M. Pavel, and H. B. Jimison are with the Consortium on Technology for Proactive Care, Northeastern University, Boston, MA 02115 USA (emails: li.xu@husky.neu.edu, i.khaghanifar@neu.edu, c.gordon@neu.edu, m.pavel@neu.edu, and h.jimison@neu.edu).

that the accuracy of wrist PPG sensors is device-specific and diminishes during movement [17-19].

One approach to mitigate the issues associated with movement-caused distortions is to examine the relationship between accuracy of the HR estimates and movement using accelerometers that are typically embedded in the monitoring devices. We operationalized movement as short-term RMS of 3-axis acceleration (RMSA) and compared heart rate data collected in a laboratory using two wrist PPG sensors (Microsoft Band 2 and E4) against a portable chest ECG sensor (Firstbeat Bodyguard 2), which we treated as a benchmark.

II. Methods

As a part of larger study, we collected physiological data from 9 healthy participants (8 males, ages 18-52). Participants were screened for mental health or heart conditions requiring medication. Subjects received $50 as remuneration for their participation. The larger study consisted of two data collection sessions in a laboratory setting and a multiple day in-the-wild data collection. In this paper, we describe our methodology for the laboratory session, which lasted approximately one hour. We collected interbeat intervals, electrodermal activity, accelerometer data from synchronized Microsoft Band 2 (MB) and E4 wrist devices with PPG sensors. We also collected interbeat intervals from Firstbeat Bodyguard 2 using ECG electrodes as our gold standard. We followed device guidelines and instructions for sensor preparation (e.g., cleaning with alcohol). Our general approach to compare the sensors was to cycle between stress and relaxation through the duration of the lab session. We first collected basic demographics (age, gender, height, weight, handedness). To understand subject trait characteristics relating to affect and stress, we administered the Perceived Stress Scale [20] and the Patient Health Questionnaire [21]. Subjects were then introduced to the devices and instrumented. The MB was worn closer to the wrist and the E4 higher up on the arm. Both wrist devices were worn clasp down with the PPG sensor facing the top of the wrist. The Firstbeat [FB] was worn on the chest according to device instructions. Participants were then led through multiple cycles of relaxation and stressor tasks to elicit physiological changes enabling us to capture a broad range values. Our stressor tasks included showing images from the International Affective Picture System (IAPS) [22] on a desktop monitor, talking aloud about the picture's contents, physical activity (marching with arms swinging and stationary biking with arms held still to compare accelerometer output), Stroop test, and mental arithmetic with distracting background noise. Relaxation tasks included sitting still in the quiet lab environment with lights off while relaxing music was played and viewing neutral IAPS imagery. Between viewing IAPS imagery and relaxation periods, participants also completed a short ecological momentary assessment (EMA) consisting of a self-assessment-manikin and a short question asking to rate their level of stress or excitement and if they had any moments that were stressful or relaxing. We chose this EMA during the lab session to provide an account of task-to-task changes in self-reported affect. Finally, a discreetly placed microphone was also used to collect the subjects' voice for future voice affect analysis.

III. Data processing

As noted in the method section, the raw output data from each sensor comprised time-stamped sequences of events representing consecutive contractions of the heart estimated either by the ECG R waves or by the PPG pulse detection. Even in the relatively controlled laboratory environment, the raw data stream contained outliers and missing samples, therefore, data cleaning was necessary (Fig. 1). There are many ways to compare the sensors that would include various metrics used to characterize HRV [23], but since the fundamental measure that is used to derive all the HRV measures is HR, our approach is based on a comparison of estimates of the HR data from each sensor.

The HR function is not directly observable and must be inferred from sensors that can detect the heart contraction either by the concomitant electric signal ECG or pressure pulse observed at various parts of the body. The ECG sensor – our gold standard – detects the QRS complex, records an estimate of the time of the R peaks of the ECG waveform, and generates a sequence of time-stamped events. Such sequences are mathematically treated as point-processes and this insight enables us to take advantage of the relevant approaches developed in mathematics. Such stochastic point processes are often described by an intensity or rate function that, in our case, corresponds to the heart rate function as a continuous representation of the sequence. Our data analyses are, therefore, based on the assumption that heart rate can be represented by a continuous intensity function $v(t)$ that describes and perhaps controls the firing of the sinoatrial node and allows a mathematical interpretation as the intensity (or rate) of the stochastic point process.

The value of the heart rate function at a given time is the consequence of a variety of factors including a competition between sympathetic and parasympathetic autonomous neural systems, breathing, etc. As such, its characteristics can then be used to estimate various aspects of the patient's state including clinical symptoms but also activity level, metabolic rate, and affective states.

Given this notion, we estimate the instantaneous heart rate function by interpolating the sequence at a uniform sampling rate that is sufficiently high to capture with an acceptable accuracy individual features of the HR function. We chose the sampling rate to be at 100 Hz. Although there are concerns involving the best way to estimate the HR function, especially for short intervals, [24], our HR estimation approach is based on a resampled version of the cubic spline interpolation – a method that has been successfully used in prior work [25].

Our approach to data cleaning is based on a model of heart rate variability where HR can be represented by a continuous band-limited function controlling the sino-atrial node of the heart, integrating complex interactions of a variety factors. This function $v(t)$ is sampled at intervals that are determined by $v(t)$ combined with internal and external noise sources. This approach is similar to integral pulse frequency modulation models [26, 27] extended to incorporate noise and variability. The objective of this work is to assess the degree of agreement between HR functions

$v(t)$ estimated from different sensors detecting RR intervals. Our initial approach is inspired by [28] but based on using robust estimates of RR variability. The RR sequence is then smoothed using singular spectrum analysis [29]. The singular spectrum analysis can be thought of as an extension of the Poincare method. We applied this process only to the data from the MB because the FB was taken as a gold standard and the E4 device already had embedded data cleaning processes.

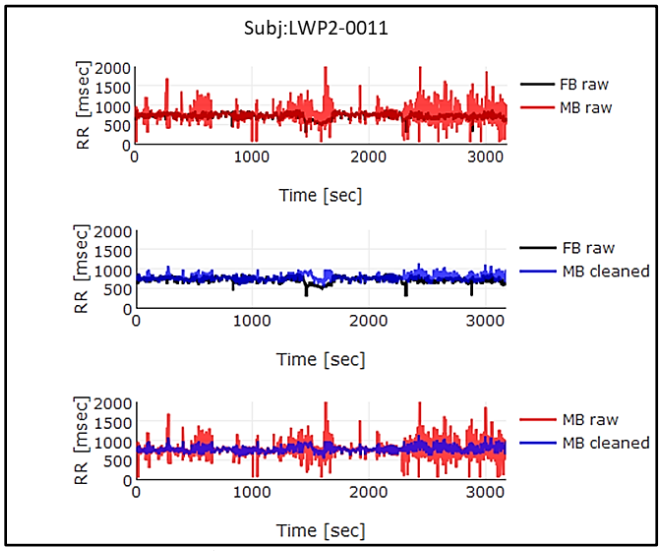

Figure 1. Comparison of raw and processed signals. The top graph is the sequence of raw RR intervals from the FB and MB. The middle one shows raw FB RR intervals and cleaned MB RR intervals. The bottom graph shows a comparison of raw and cleaned MB data.

IV. Statistical Analysis

In our statistical analyses, we used the cleaned data (as described above) unless otherwise stated. Accelerometry signals from E4 and MB were upsampled to 100 Hz. Next, all signals, including RR-intervals, were downsampled to 20 Hz commensurate with the frequency content of the signals. Gravity was removed from each axis individually by computing a mean of a rolling window of width 1 minute for a given axis and subtracting this mean from each sample for this axis. We used resulting values to compute RMS of acceleration. RR signals were aligned using a lag obtained from cross-correlation.

The next step involved a selection of a metric for comparisons of the HR functions from different sensors. Following previous research [30], we used the concordance correlation coefficient [31], which we computed over a 1-minute rolling window, because it accounts for systematic bias in HR and the lack of linear correlation.

A representative example of the resulting correlation between the FB and E4 & MB data computed for the period of the laboratory study is shown in Fig. 2. The top graph represents the correlation for each window temporal position; the bottom graph represents the RMS of acceleration computed over the same window.

We modeled the relationship between agreement of RR signals and motion using a linear regression with RMS of acceleration (RMSA) as predictor of concordance correlation coefficient (CCC). To assess performance of our data cleaning approach, we computed two CCCs and fitted two models; one using raw data and one with data cleaned using singular spectrum analysis. Our approach improved CCC of MB and FB (Fig.3). Mean of $CCC_{CLEANED}$ (.324) is statistically significantly higher than CCC_{RAW} (.283) (Welch Two Sample t-test, t = 19.764, p < .001). Furthermore, regression analysis revealed that the impact of RMSA on CCC is smaller when data is smoothed. ($\beta_{RMSA\ CLEANED}$ = -.109, $p_{RMSA\ CLEANED}$ < .001, R^2 = .003; $\beta_{RMSA\ RAW}$ = -.942, $p_{RMSA\ RAW}$ < .001, R^2 = .236). These results indicate that our approach improves HR estimates; some of the corrected artifacts seem to be originating from motion artifacts.

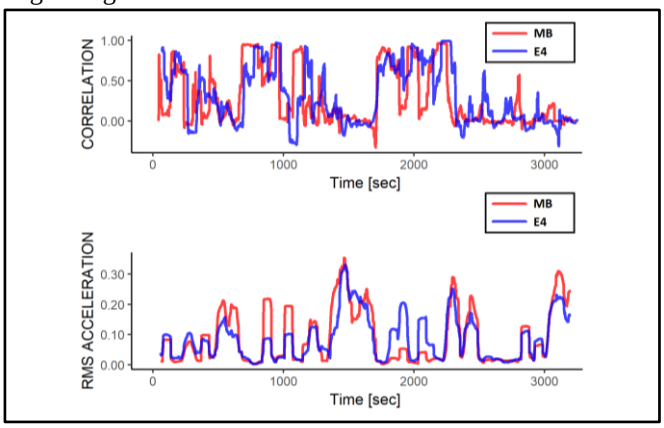

Figure 2. Example of correlation between FB and MB & E4 sensors. Gains of acceleration were adjusted.

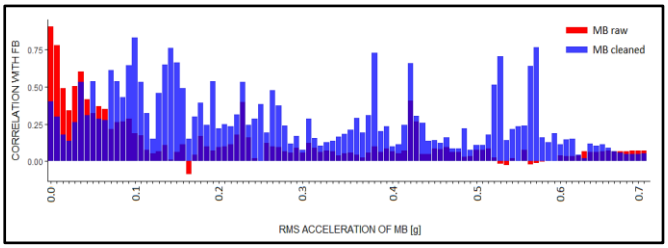

Figure 3. Correlation between FB and MB raw & MB cleaned RR signals. Bars represent means of correlation computed over binned values of RMSA. Cleaned MB data (blue) is more highly correlated with FB than raw MB data (red) for most values of RMSA.

To estimate the impact of RMSA on CCC we fitted two models using not cleaned data, one for MB and one for E4. For each sensor, the relationship between RMSA and CCC was statistically significant. RMSA was statistically significant predictors of CCC between FB and MB (β_{RMSA} = -.942, p_{RMSA} < .001, R^2 = .236). Similarly, RMSA was a statistically significant predictor of CCC between FB and E4 (β_{RMSA} = -1.234, p_{RMSA} < .001, R^2 = .066). The mean of CCC_{E4} (.300) is statistically significantly higher than CCC_{MB} (.283) (Welch Two Sample t-test, t = 8.698, p < .001).

V. Results

Preliminary analysis of data from the whole study showed that the results from the data of one subject presented here are representative of the general results across all nine participants. Our analysis demonstrated that the coherence between the PPG and the ECG data is not as high as one would like to assure accurate HR assessment. However, our

novel data cleaning approach can be used to improve coherence. We also detected that RMSA has a large, negative impact on the agreement between PPG and ECG signals. Researchers relying on HRV estimates obtained from PPG sensors may consider using RMSA to assess the degree to which they should trust these estimates. There is a good chance that a more fine-grained analysis, perhaps relying on nonlinear methods and fusion of the accelerometry with the HR estimate, might be used to further correct motion-based distortions.

VI. CONCLUSION

Our study has indicated the need for algorithms using RMSA to determine when HRV estimates from a wrist device are sufficiently accurate for health coaching feedback. We propose including the above factors as parameters in modeling sensor agreement with the gold standard ECG signal. The promise of using wrist PPG sensors to monitor HRV in real time offers many opportunities to improve continuous health coaching interventions. Although the accuracy for these real time estimates need not be perfect, knowing when they are reliable can offer a substantial improvement to the effectiveness of the coaching interventions.

VII. REFERENCES

[1] F. S. Collins and H. Varmus, "A new initiative on Precision Medicine," *New England Journal of Medicine,* vol. 372, pp. 793-795, 2015.

[2] K. Hamill. (2017, April 28). *Worldwide Efforts to Accelerate Precision Medicine*. Available: https://sciex.com/community/blogs/blogs/worldwide-efforts-to-accelerate-precision-medicine

[3] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, *et al.*, "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support," *Annals of Behavioral Medicine*, pp. 1-17, 2016.

[4] R. E. Kleiger, P. K. Stein, and J. T. Bigger, "Heart rate variability: measurement and clinical utility," *Annals of Noninvasive Electrocardiology,* vol. 10, pp. 88-101, 2005.

[5] P. Ponikowski, M. Piepoli, T. Chua, W. Banasiak, D. Francis, S. Anker, *et al.*, "The impact of cachexia on cardiorespiratory reflex control in chronic heart failure," *European heart journal,* vol. 20, pp. 1667-1675, 1999.

[6] J. S. Wu, Y. C. Yang, F. H. Lu, T. S. Lin, J. J. Chen, Y. H. Huang, *et al.*, "Cardiac autonomic function and insulin resistance for the development of hypertension: A six-year epidemiological follow-up study," *Nutrition, Metabolism and Cardiovascular Diseases,* vol. 23, pp. 1216-1222, 12// 2013.

[7] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers Iii, and T. D. Wager, "A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health," *Neuroscience & Biobehavioral Reviews,* vol. 36, pp. 747-756, 2// 2012.

[8] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Review of general psychology,* vol. 10, p. 229, 2006.

[9] S. C. Segerstrom and G. E. Miller, "Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry," *Psychological bulletin,* vol. 130, p. 601, 2004.

[10] "Mental health action plan 2013-2020," W. H. Organization, Ed., ed, 2013.

[11] H. V. Huikuri, T. H. Mäkikallio, and J. Perkiömäki, "Measurement of heart rate variability by methods based on nonlinear dynamics," *Journal of electrocardiology,* vol. 36, pp. 95-99, 2003.

[12] G. Ernst, "Methodological Issues," in *Heart Rate Variability*, ed London: Springer London, 2014, pp. 51-118.

[13] C.-W. Lin, J.-S. Wang, and P.-C. Chung, "Mining physiological conditions from heart rate variability analysis," *IEEE Computational Intelligence Magazine,* vol. 5, pp. 50-58, 2010.

[14] E. Gil, M. Orini, R. Bailón, J. Vergara, L. Mainardi, and P. Laguna, "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiological measurement,* vol. 31, p. 1271, 2010.

[15] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International Journal of Cardiology,* vol. 166, pp. 15-29, 6/5/ 2013.

[16] K. R. Evanson, M. M. Goto, and R. D. Furberg, "Systematic review of the validity and reliability of consumer-wearable activity trackers," *International Journal of Behavioral Nutrition and Physical Activity,* vol. 12, 2015.

[17] K. Lewis, "Validation of Wearable Biofeedback Technology for Heart Rate Tracking Via Reflective Photoplethymorgraphy," California State Polytechnic University, Pomona, 2017.

[18] J. Parak and I. Korhonen, "Evaluation of wearable consumer heart rate monitors based on photopletysmography," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2014, pp. 3670-3673.

[19] M. P. Wallen, S. R. Gomersall, S. E. Keating, U. Wisløff, and J. S. Coombes, "Accuracy of heart rate watches: Implications for weight management," *PloS one,* vol. 11, p. e0154420, 2016.

[20] S. Levenstein, C. Prantera, V. Varvo, M. L. Scribano, E. Berto, C. Luzi, *et al.*, "Development of the perceived stress questionnaire: A new tool for psychosomatic research," *Journal of Psychosomatic Research*, vol. 37, pp. 19-32, 1993.

[21] A. Martin, W. Rief, A. Klaiberg, and E. Braehler, "Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population," *General hospital psychiatry,* vol. 28, pp. 71-77, 2006.

[22] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Instruction manual and affective ratings," *The center for research in psychophysiology, University of Florida*, 1999.

[23] T. F. o. t. E. S. o. Cardiology, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *Eur heart J,* vol. 17, pp. 354-381, 1996.

[24] T. Thong, K. Li, J. McNames, M. Aboy, and B. Goldstein, "Accuracy of ultra-short heart rate variability measures," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, 2003, pp. 2424-2427.

[25] D.-G. Jang, M. Hahn, J.-K. Jang, U. Farooq, and S.-H. Park, "A comparison of interpolation techniques for RR interval fitting in AR spectrum estimation," in *Biomedical Circuits and Systems Conference (BioCAS), 2012 IEEE*, 2012, pp. 352-355.

[26] J. Mateo and P. Laguna, "New heart rate variability time-domain signal construction from the beat occurrence time and the IPFM model," in *Computers in Cardiology, 1996*, 1996, pp. 185-188.

[27] J. Mateo and P. Laguna, "Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal," *IEEE Transactions on Biomedical Engineering,* vol. 50, pp. 334-343, 2003.

[28] G. G. Berntson, K. S. Quigley, J. F. Jang, and S. T. Boysen, "An approach to artifact identification: Application to heart period data," *Psychophysiology,* vol. 27, pp. 586-598, 1990.

[29] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for time series*: Springer Science & Business Media, 2013.

[30] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International Journal of Cardiology,* vol. 166, pp. 15-29, 2013.

[31] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255-268, 1989.